Projet 5 : Segmentez des clients d'un site e-commerce

30/08/2024

Soukaina GUAOUA ELJADDI

Parcours Data Scientist
OpenClassrooms

Plan:

- Problématique et présentation du jeu de données
- Analyse exploratoire des données et feature engineering
- Essais de différents approches de Modélisation
- Simulation d’un contrat de maintenance
- Conclusion

Problématique et présentation du jeu de données

Problématique

Contexte : Entreprise brésilienne '**Olist**' qui propose une **solution de vente** sur les **marketplaces en ligne**, et qui veut monter une équipe data avec un 1er projet sur la **segmentation client**.

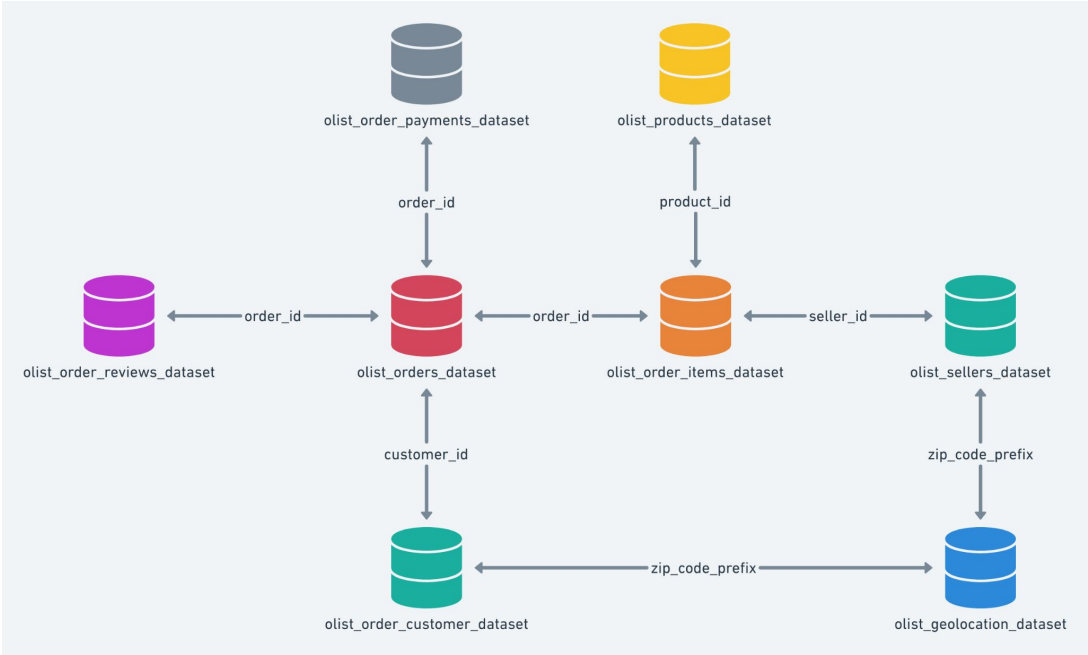
Objectifs et missions :

- Fournir aux équipes d'e-commerce une s**egmentation des clients** pour les **campagnes de communication**
- Comprendre les différents **types d'utilisateurs**
- Fournir une **description actionable de la segmentation**
- Faire une proposition de **contrat de maintenance**

Présentation du jeu de données

Base de données '**Olist**' avec **9 fichiers CSV** (132 Mo)

	name
0	customers
1	geoloc
2	order_items
3	order_pymts
4	order_reviews
5	orders
6	products
7	sellers
8	translation



Présentation du jeu de données

Base de données '**Olist**' :

- une **BD** gratuite, anonymisée mise en ligne sur **Kaggle**
- des **données variées** (textuelles, chiffrées, catégorielles, géographiques)
- **données commerciales** Olist sur 2 ans, de **2016** à **2018**
- **96 096** clients uniques concernés
- **99 441** commandes distinctes

Partie 1 : Analyse exploratoire des données et feature engineering

Partie 1 : **Analyse exploratoire des données**

```
# Afficher les données sur customers
customers.head()
```

	index	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
0	0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
1	1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
2	2	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP
3	3	b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP
4	4	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP

Partie 1 : **Analyse exploratoire des données**

```
# Afficher des informations générales sur customers
customers.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   index                   99441 non-null  int64
 1   customer_id             99441 non-null  object
 2   customer_unique_id      99441 non-null  object
 3   customer_zip_code_prefix  99441 non-null  int64
 4   customer_city           99441 non-null  object
 5   customer_state          99441 non-null  object
dtypes: int64(2), object(4)
memory usage: 4.6+ MB
```

Partie 1 : **Analyse exploratoire des données**

```
# Statistiques descriptives
customers.describe()
```

	index	customer_zip_code_prefix
count	99441.000000	99441.000000
mean	49720.000000	35137.474583
std	28706.288396	29797.938996
min	0.000000	1003.000000
25%	24860.000000	11347.000000
50%	49720.000000	24416.000000
75%	74580.000000	58900.000000
max	99440.000000	99990.000000

Partie 1 : **Analyse exploratoire des données**

```
# Analyse par catégorie de produit
product_categories = products.groupby('product_category_name')['product_id'].count()
product_categories.name = 'product_count'
product_categories
```

```
product_category_name
agro_industria_e_comercio      74
alimentos                      82
alimentos_bebidas             104
artes                          55
artes_e_artesanato             19
                              ...
sinalizacao_e_seguranca        93
tablets_impressao_imagem        9
telefonia                    1134
telefonia_fixa                116
utilidades_domesticas        2335
Name: product_count, Length: 73, dtype: int64
```

Partie 1 : **Analyse exploratoire des données**

	Date de commande	Date estimé de livraison	Date de livraison	Délai de livraison	Date de review
0	2017-10-02	2017-10-18	2017-10-10	-8.0	2018-01-18
1	2018-07-24	2018-08-13	2018-08-07	-6.0	2018-03-10
2	2018-08-08	2018-09-04	2018-08-17	-18.0	2018-02-17
3	2017-11-18	2017-12-15	2017-12-02	-13.0	2017-04-21
4	2018-02-13	2018-02-26	2018-02-16	-10.0	2018-03-01

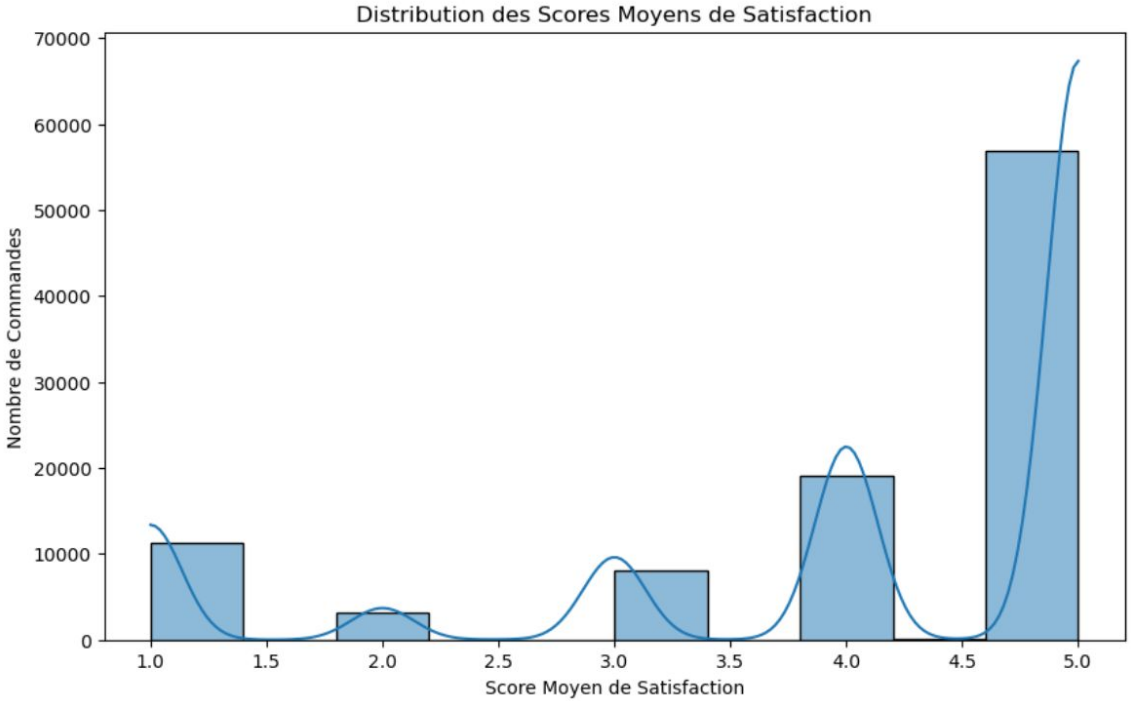
Partie 1 : **Analyse exploratoire des données**

```
# Analyse de la satisfaction client
review_scores = order_reviews.groupby('order_id')['review_score'].mean()
review_scores.name = 'avg_review_score'
review_scores
```

```
order_id
00010242fe8c5a6d1ba2dd792cb16214    5.0
00018f77f2f0320c557190d7a144bdd3    4.0
000229ec398224ef6ca0657da4fc703e    5.0
00024acbcdf0a6daa1e931b038114c75    4.0
00042b26cf59d7ce69dfabb4e55b4fd9    5.0
                                   ...
fffc94f6ce00a00581880bf54a75a037    5.0
fffcd46ef2263f404302a634eb57f7eb    5.0
fffce4705a9662cd70adb13d4a31832d    5.0
fffe18544ffabc95dfada21779c9644f    5.0
fffe41c64501cc87c801fd61db3f6244    5.0
Name: avg_review_score, Length: 98673, dtype: float64
```

Partie 1 : Analyse exploratoire des données

Satisfaction client:



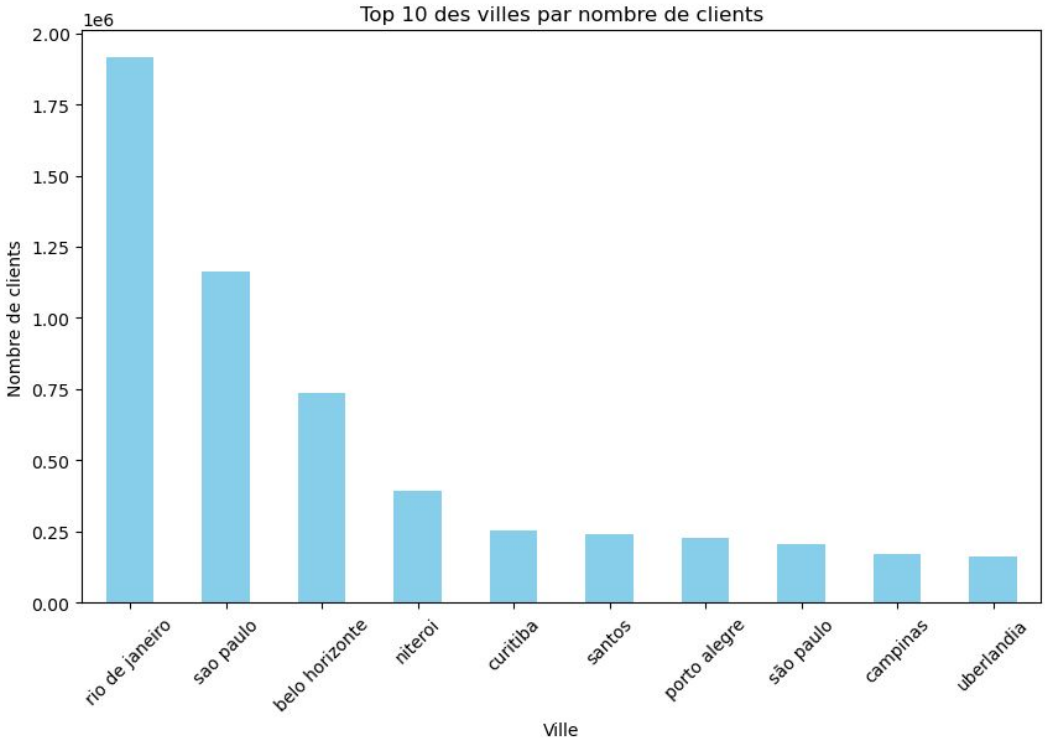
Partie 1 : Analyse exploratoire des données

Tableau de bord des performances des vendeurs :

seller_id	price	review_score	delay_days
0015a82c2db000af6aaaf3ae2ecb0532	2685.00	3.66	-16.33
001cca7ae9ae17fb1caed9dfb1094831	25080.03	3.90	-13.21
001e6ad469a905060d959994f1b41e4f	250.00	1.00	NaN
002100f778ceb8431b7a1020ff7ab48f	1254.40	3.98	-8.21
003554e2dce176b5555353e4f3555ac8	120.00	5.00	-27.00

Partie 1 : Analyse exploratoire des données

La distribution des clients par ville (top 10)



Partie 1 : **Feature engineering**

La méthode **RFM** (Récence, Fréquence, Montant) pour le ciblage marketing.

Récence (Recency) : Nombre de jours depuis la dernière commande de chaque client.

Fréquence (Frequency) : Nombre total de commandes passées par chaque client.

Montant (Monetary) : Montant total dépensé par chaque client.

Partie 1 : **Feature engineering**

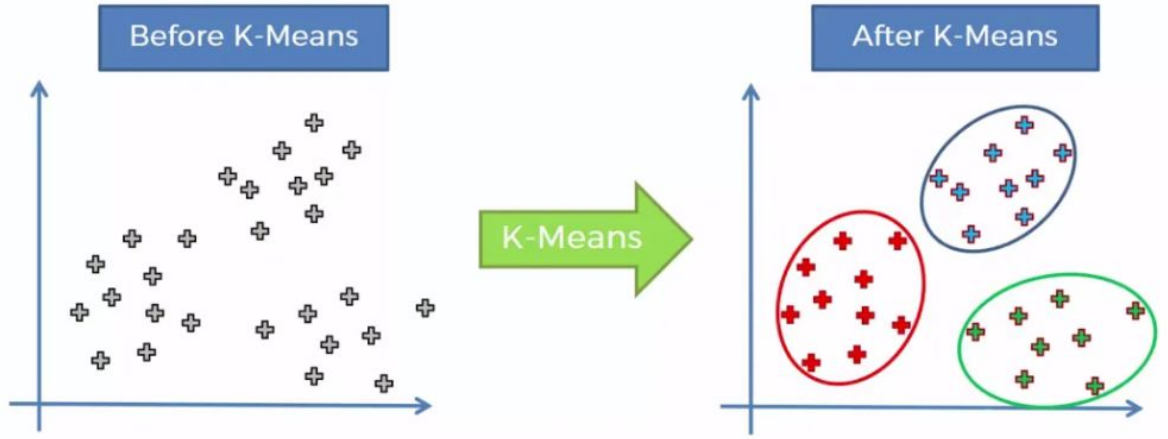
RFM :

	customer_id	recency	frequency	monetary
0	00012a2ce6f8dcda20d059ce98491703	337.056852	1	89.80
1	000161a058600d5901f007fab4c27140	458.326227	1	54.90
2	0001fd6190edaaf884bcaf3d49edf079	596.266377	1	179.99
3	0002414f95344307404f0ace7a26f1d5	427.181227	1	149.90
4	000379cdec625522490c315e70c7a9fb	198.158345	1	93.00

Partie 2 : Essais de différents approches de Modélisation

Partie 2 :Essais de différents approches de Modélisation

K-means: un algorithme d'apprentissage non supervisé, qui permet d'analyser un jeu de données afin de regrouper les données « similaires » en groupes (ou clusters)



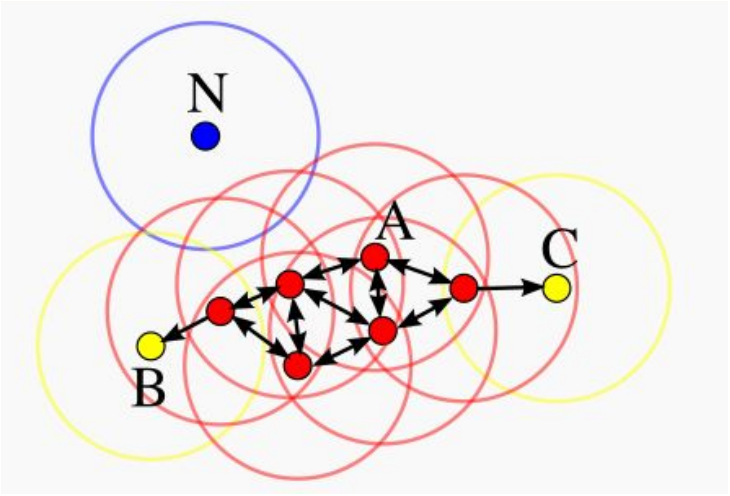
medium.datadriveninvestor.com

Partie 2 : Essais de différents approches de Modélisation

DBSCAN : un algorithme de clustering fondé sur la densité, il identifie les régions denses de points et les considère comme des clusters.

Paramètres Clés :

- **Epsilon (ε\varepsilonε)** : La distance maximale entre deux points pour qu'ils soient considérés comme voisins.
- **MinPts (Minimum Points)** : Le nombre minimum de points requis pour former un cluster dense.



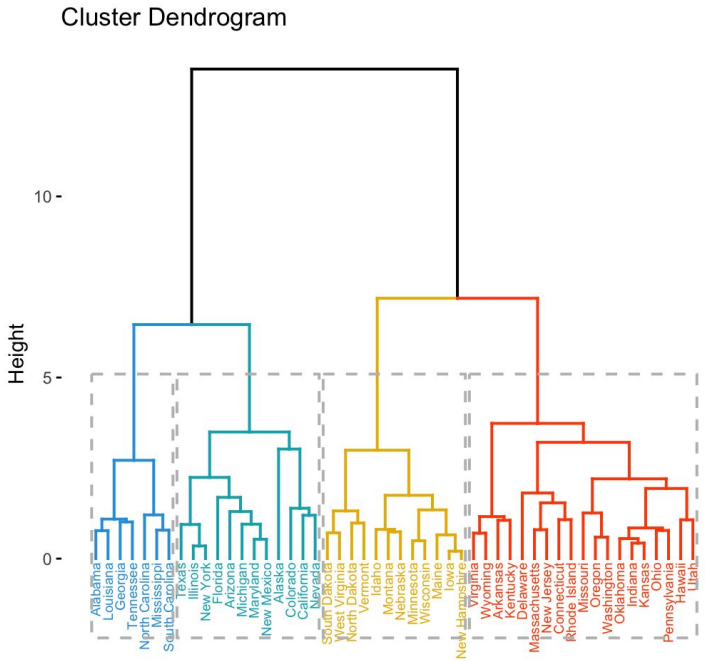
wikipedia

Partie 2 : **Essais de différents approches de Modélisation**

CAH ou la **Clustering Hiérarchique Ascendant** : méthode de regroupement qui construit une hiérarchie de clusters en procédant par étapes successives.

Elle ne nécessite pas de spécifier le nombre de clusters à l'avance.

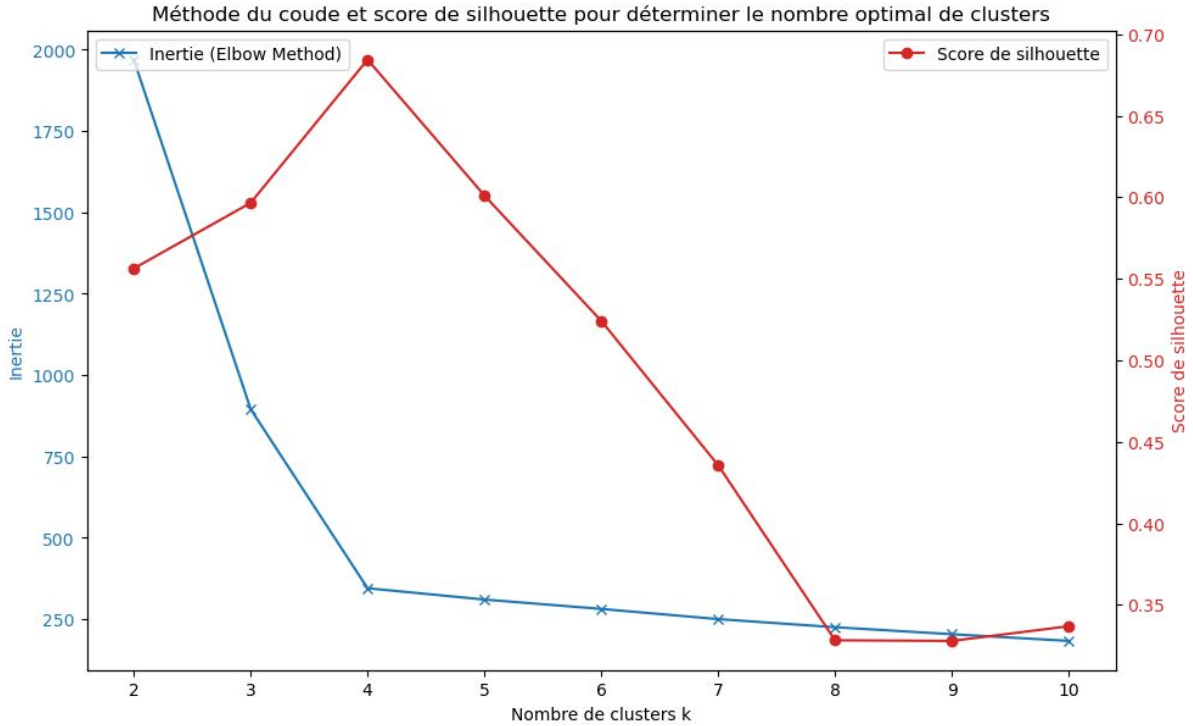
La CAH construit un arbre de clusters (ou dendrogramme) en fusionnant ou en divisant des groupes de données successivement.



https://www.imo.universite-paris-saclay.fr/

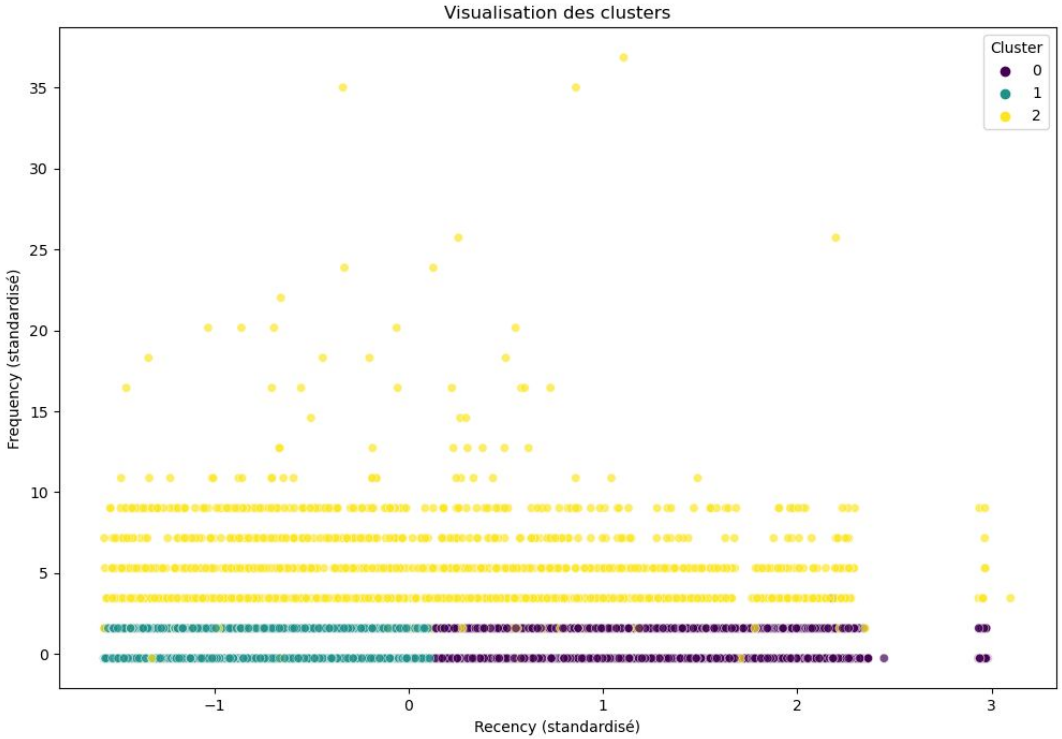
Partie 2 : Essais de différents approches de Modélisation

K-means : détermination du nombre de clusters optimal



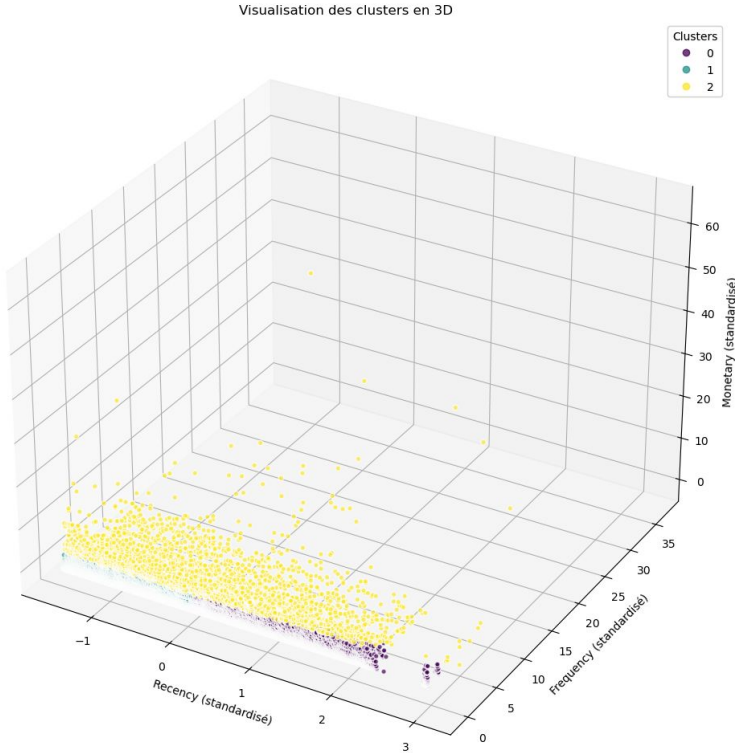
Partie 2 : Essais de différents approches de Modélisation

K-means : visualisation des clusters 2D



Partie 2 : Essais de différents approches de Modélisation

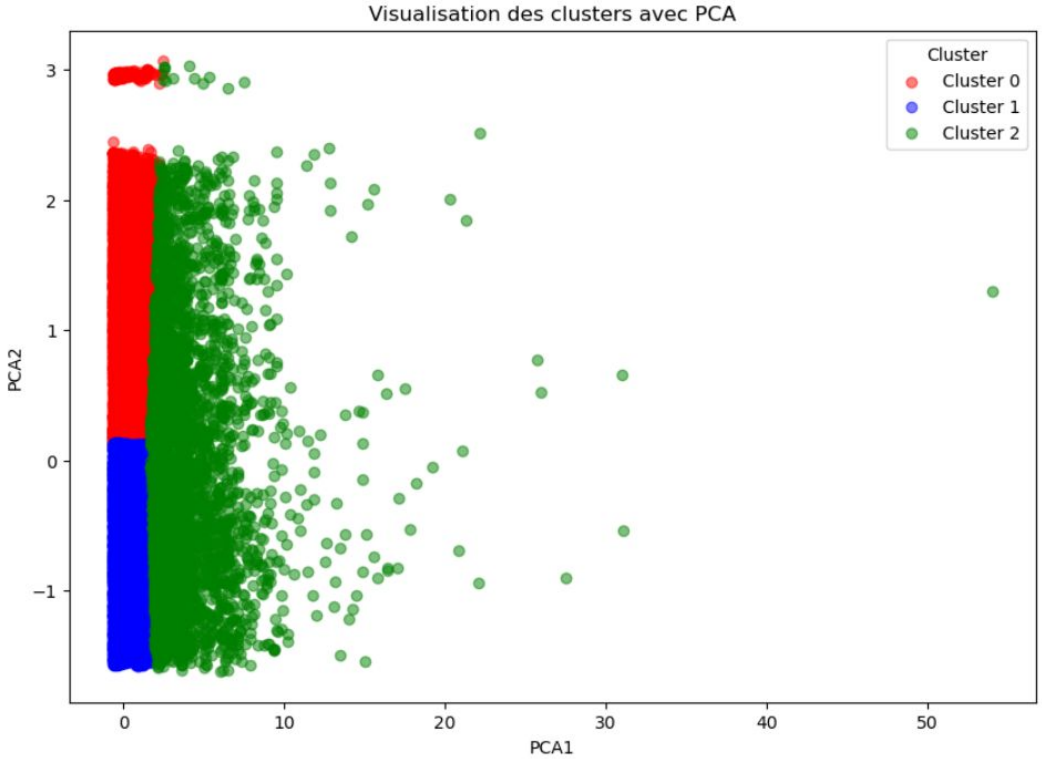
K-means : visualisation des clusters 3D



Partie 2 : **Essais de différents approches de Modélisation**

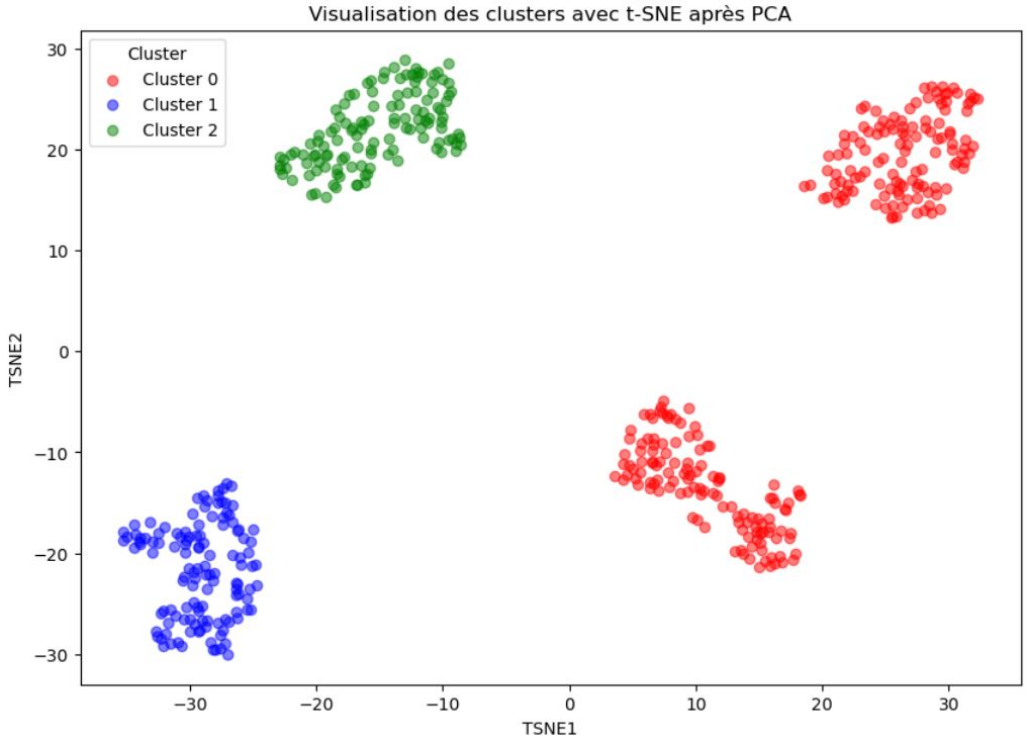
K-means : visualisation

des clusters après ACP



Partie 2 : Essais de différents approches de Modélisation

K-means : visualisation des clusters avec t-sne



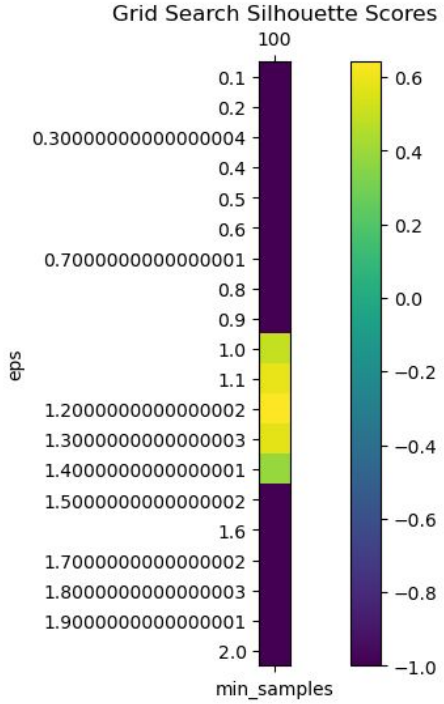
Partie 2 : Essais de différents approches de Modélisation

DBSCAN : détermination de l'eps et min_samples optimales

Best eps: 1.2,

Best min_samples: 100,

Best silhouette score: 0.64



Partie 2 : **Essais de différents approches de Modélisation**

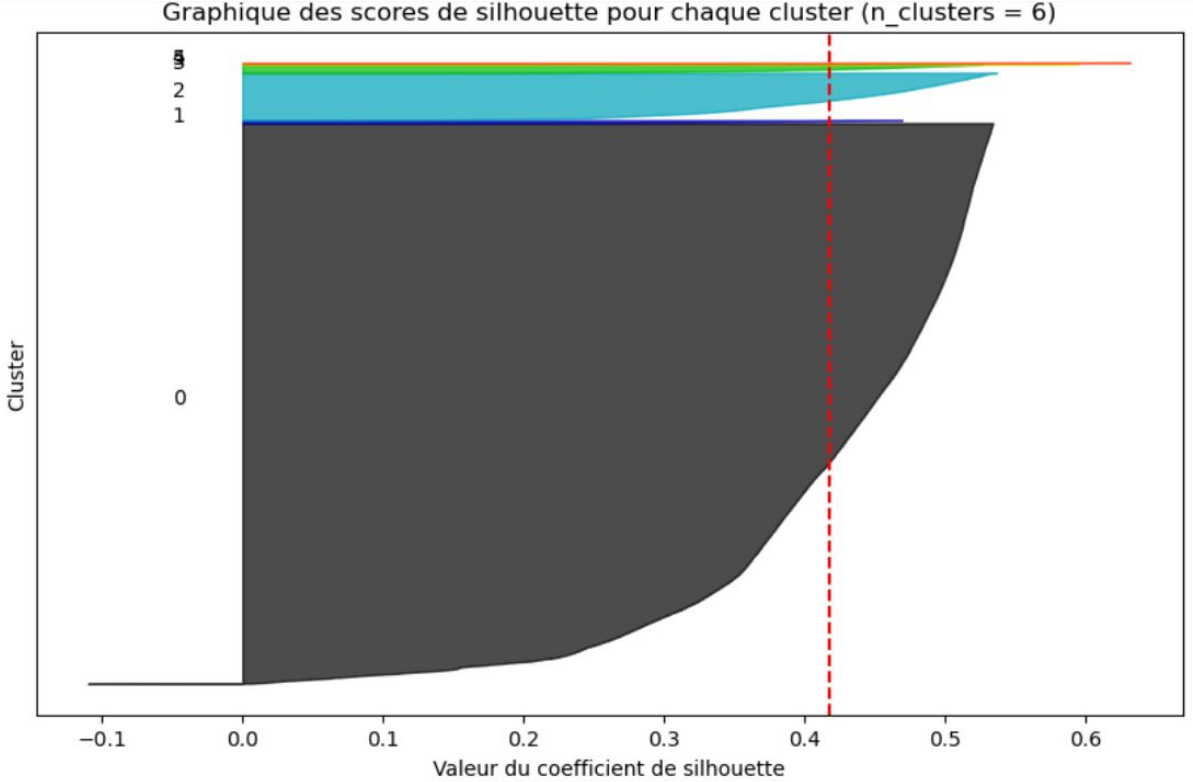
DBSCAN:

évaluation

de la qualité

de clustering

avec DBSCAN



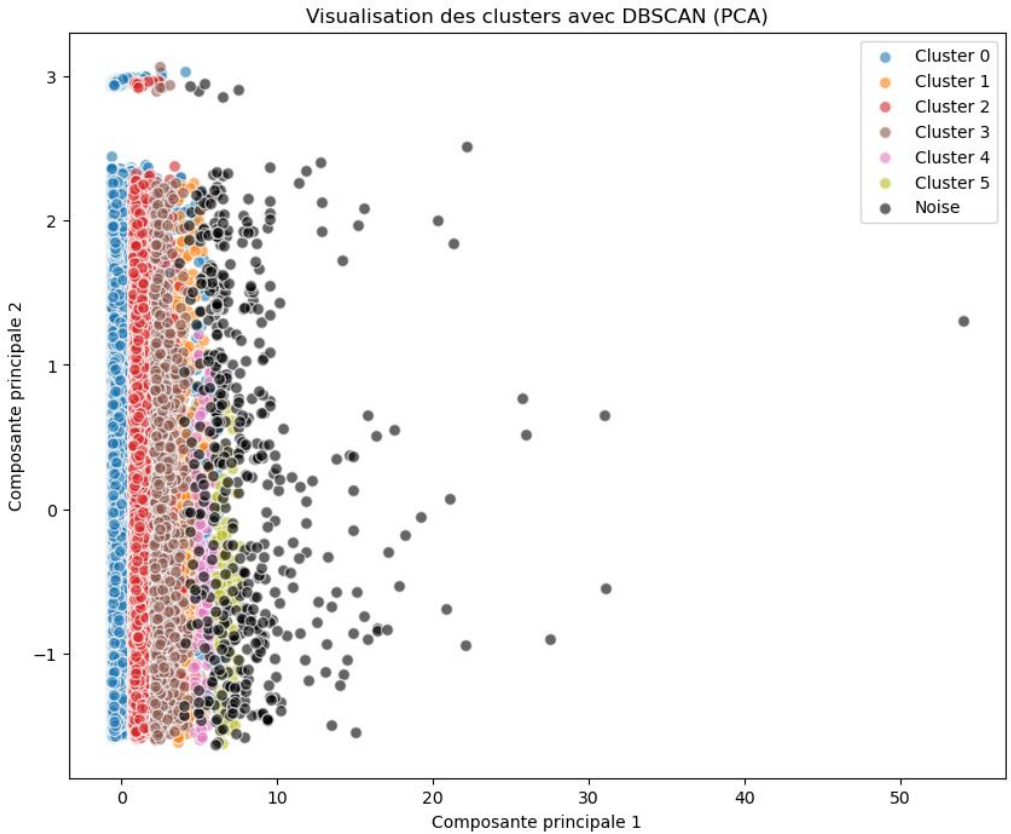
Partie 2 : **Essais de différents approches de Modélisation**

DBSCAN:

Visualisation des

6 clusters obtenus

en utilisant l'ACP



Partie 2 : **Essais de différents approches de Modélisation**

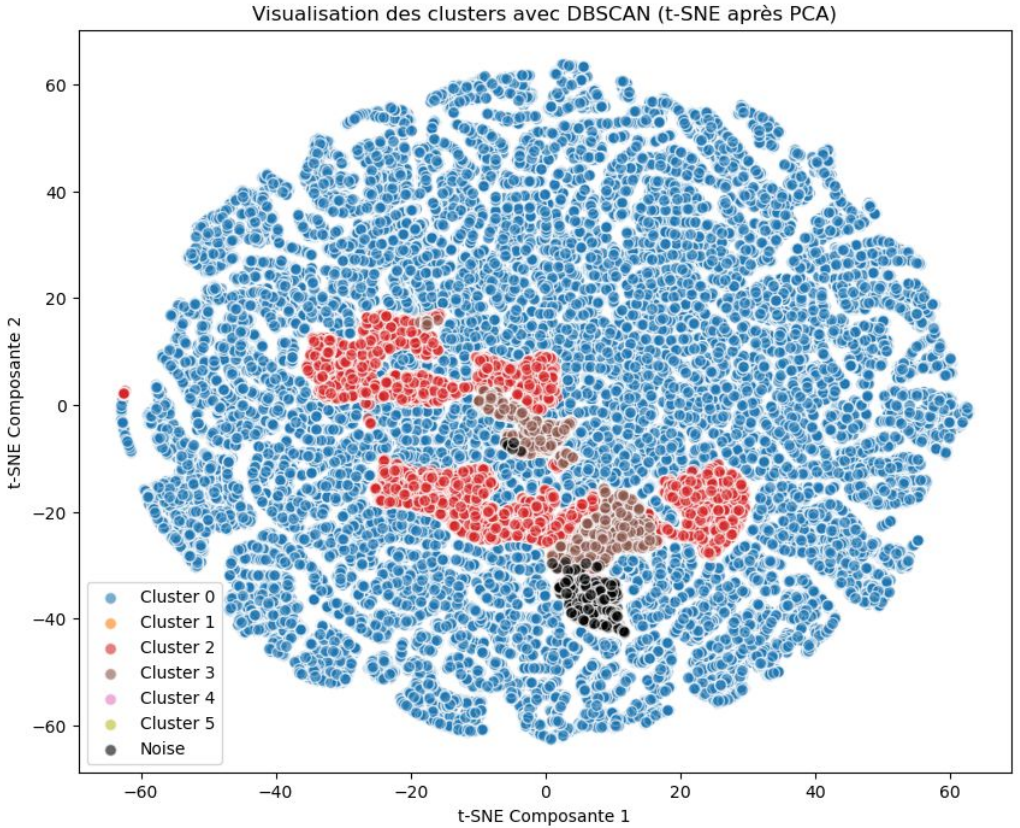
DBSCAN:

Visualisation des

6 clusters obtenus

en utilisant

l'ACP + t-sne

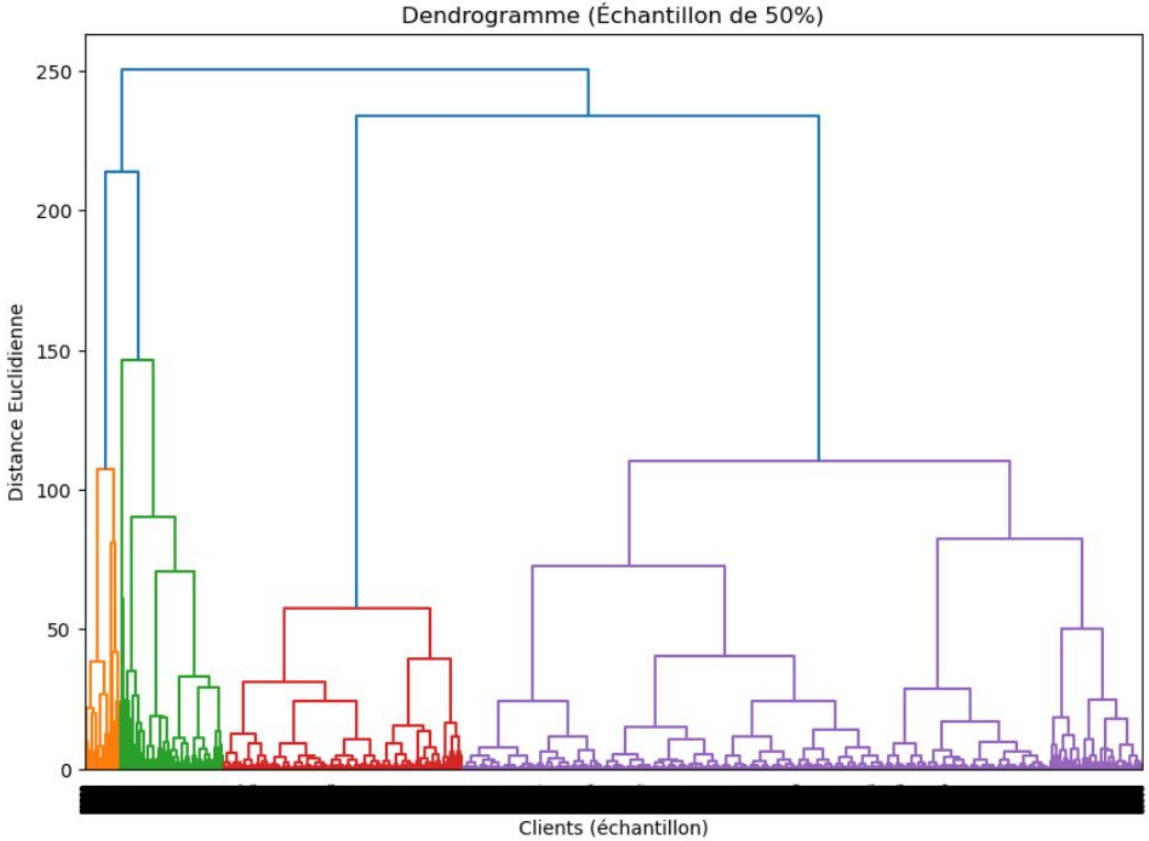


Partie 2 : **Essais de différents approches de Modélisation**

CAH :

5 clusters

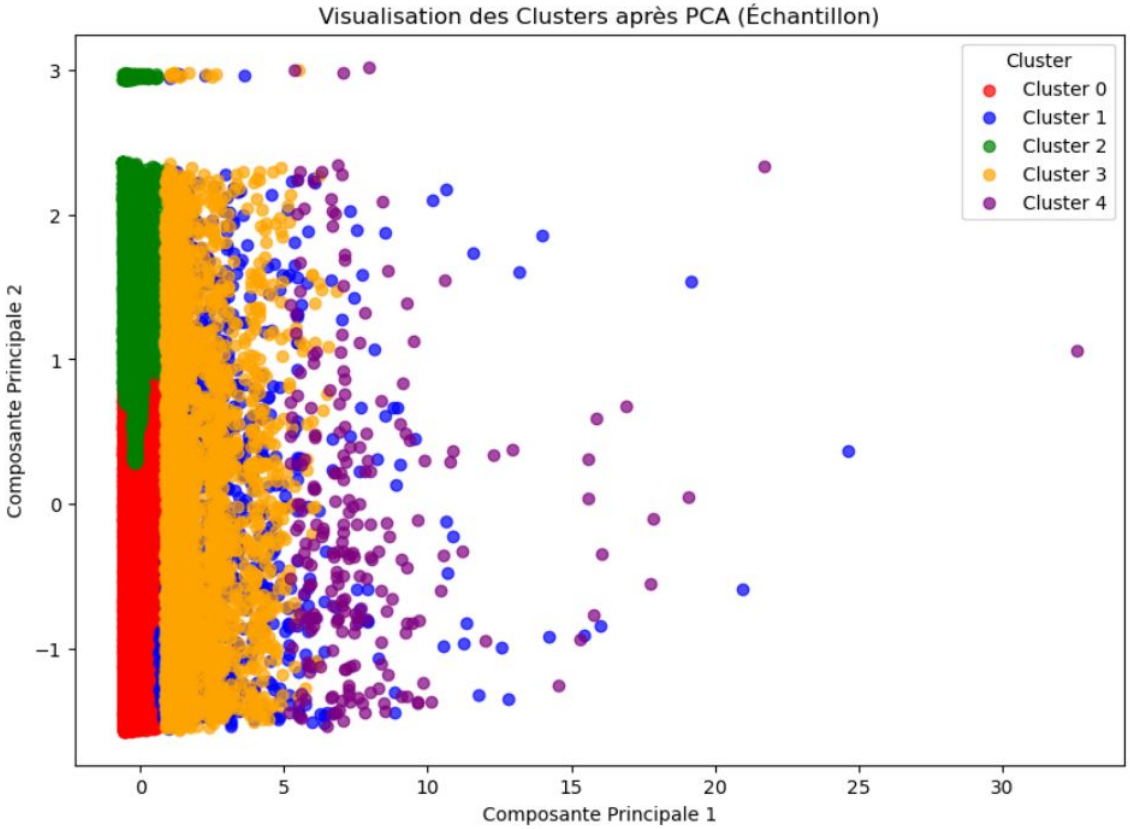
identifiés



Partie 2 : Essais de différents approches de Modélisation

CAH :

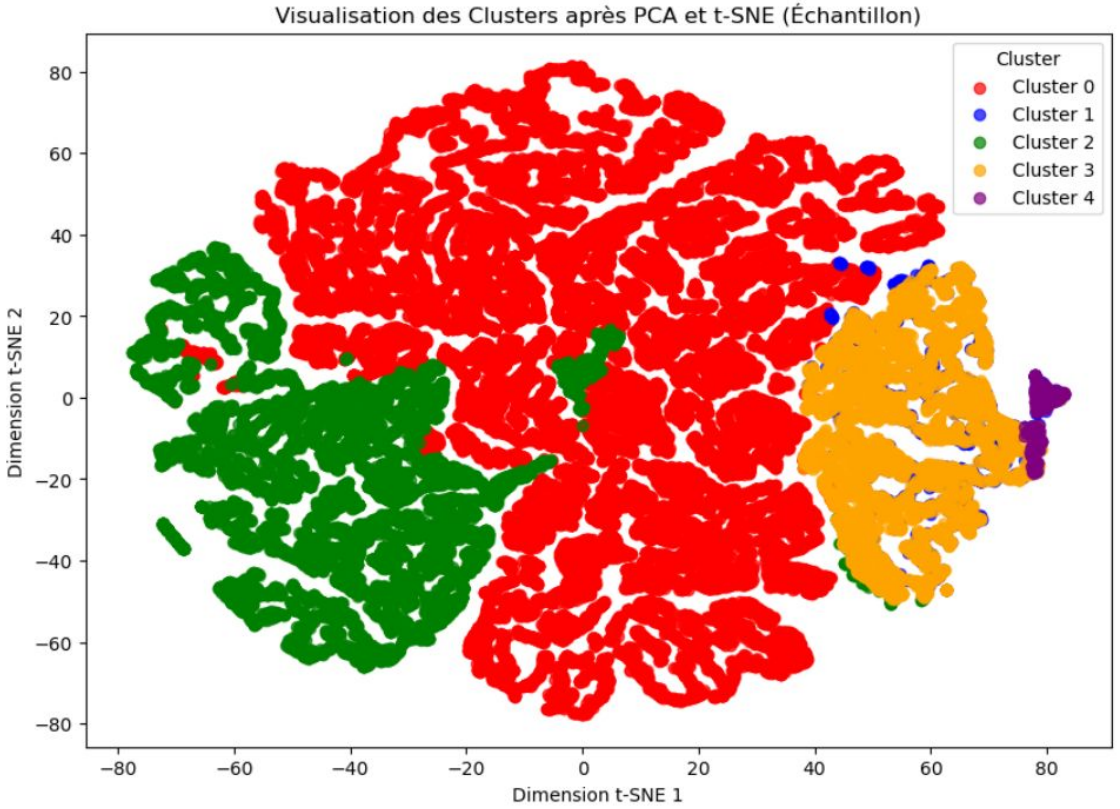
Visualisation des clusters obtenus avec l'ACP



Partie 2 : Essais de différents approches de Modélisation

CAH :

Visualisation des clusters obtenus avec l'ACP + t-sne



Partie 2 : Essais de différents approches de Modélisation

Evaluation de la qualité du clustering avec les métriques :

1. Score de silhouette

- **Quoi** : Évalue la cohésion (proximité des points au sein d'un même cluster) et la séparation (distance entre les clusters).
- **Valeurs** :
  - Proche de **1** = Bon clustering (points bien regroupés et clusters bien séparés).
  - Proche de **0** = Points proches des frontières entre clusters.
  - Négatif = Mauvaise affectation des points.

Partie 2 : Essais de différents approches de Modélisation

Evaluation de la qualité du clustering avec les métriques :

2. Coefficient de Davies-Bouldin

- **Quoi** : Mesure la moyenne des pires ratios de similarité entre clusters. Combinaison de compacité et séparation.
- **Valeurs** :
  - Plus bas = Meilleure séparation et compacité.

Partie 2 : Essais de différents approches de Modélisation

Evaluation de la qualité du clustering avec les métriques :

3. Index de Calinski-Harabasz

- **Quoi** : Rapport entre la variance inter-cluster (séparation) et intra-cluster (compacité).
- **Valeurs** :
  - Plus élevé = Clusters bien séparés et compacts.

Partie 2 : Essais de différents approches de Modélisation

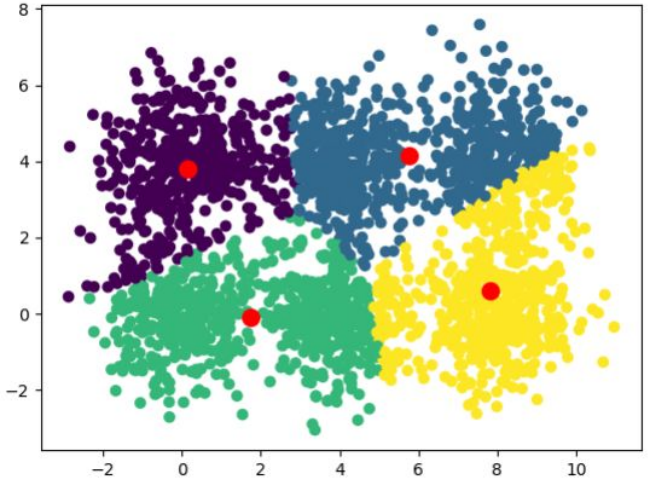
Comparaison entre les 3 méthodes de clustering après ACP:

	K-means	DBSCAN	CAH
Score de silhouette	0.478	0.323	0.463
Coefficient de Davies-Bouldin	0.727	1.291	0.766
Index de Calinski-Harabasz	76435.975	6045.697	22074.959

Partie 2 : Essais de différents approches de Modélisation

Interprétation des clusters d'un point de vue métier: analyse des centroïdes des clusters

Les **centroïdes** des clusters sont des points qui représentent la "moyenne" ou le centre des points appartenant à chaque cluster dans un espace de caractéristiques.



khayyam.dev eloppez.com/

$$\text{Centroïde}_i = \frac{1}{N} \sum_{j=1}^{N} X_{ij}$$

où Xij est la valeur de la caractéristique i du point j, et N est le nombre de points dans le cluster.

Partie 2 : Essais de différents approches de Modélisation

Interprétation des clusters d'un point de vue métier: analyse des centroïdes des clusters

clusters	recency	frequency	monetary
0	0.976706	-0.123726	-0.103287
1	-0.723421	-0.130162	-0.109941
2	-0.033731	2.963620	2.491135

Recency : Nombre de jours depuis la dernière interaction ou achat du client.

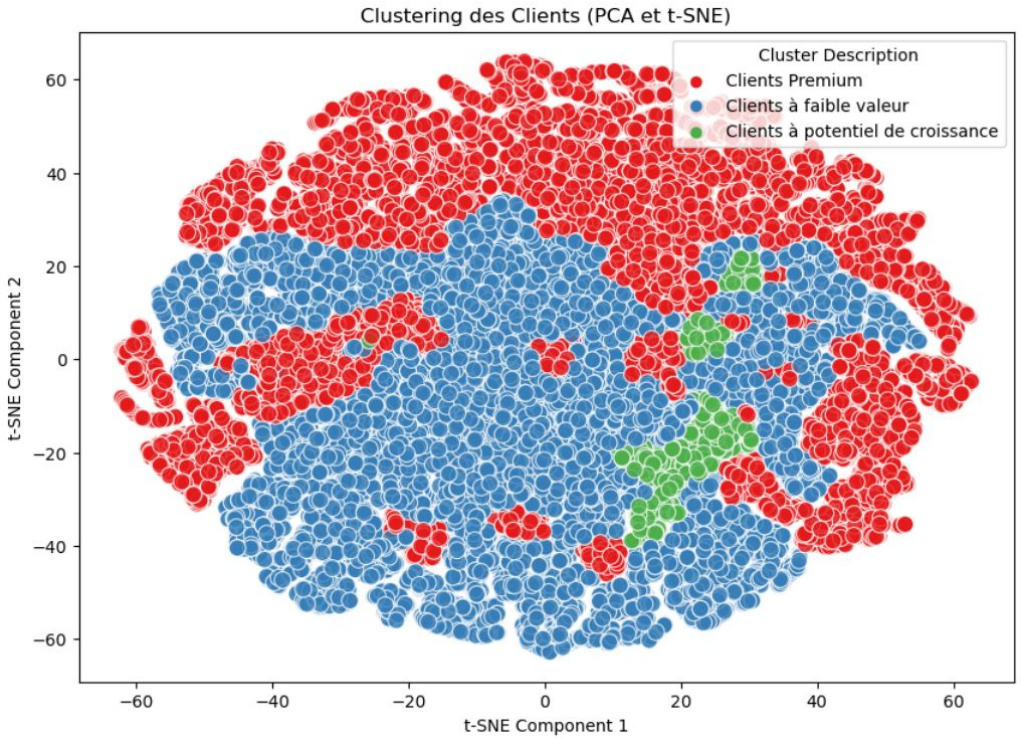
Frequency : Nombre total d'achats ou d'interactions effectués par le client.

Monetary : Montant total dépensé par le client.

Partie 2 : **Essais de différents approches de Modélisation**

Visualisation des clusters

‘clients’ avec ACP + t-sne



Partie 2 : Essais de différents approches de Modélisation

Interprétation des Centroïdes et des Descriptions de Clusters

Cluster 0 : **Clients Premium**

Recency : 0.976706 (élevée) -> Ces clients ont une valeur élevée sur l'échelle de recency, ce qui signifie qu'ils ont acheté récemment.

Frequency : -0.123726 (légèrement négative) -> Leur fréquence d'achat n'est pas particulièrement élevée.

Monetary : -0.103287 (légèrement négative) -> Ils ne dépensent pas énormément.

Interprétation : Ces clients ne sont pas des acheteurs fréquents ni de gros dépensiers. Mais leur récente activité pourrait indiquer une tendance à devenir plus actifs. Ils sont qualifiés de "Clients Premium" en raison de leur récente activité, suggérant qu'ils sont engagés ou potentiellement intéressés par de nouveaux achats.

Partie 2 : **Essais de différents approches de Modélisation**

Interprétation des Centroïdes et des Descriptions de Clusters

Cluster 1 : **Clients à faible valeur**

Recency : -0.723421 (faible) -> Ces clients ont acheté il y a longtemps.

Frequency : -0.130162 (légèrement négative) -> Ils n'achètent pas souvent.

Monetary : -0.109941 (légèrement négative) -> Ils dépensent peu.

Interprétation : Ces clients sont moins actifs, achètent rarement, et dépensent peu lorsqu'ils le font. Ils peuvent être considérés comme des clients à faible valeur pour l'entreprise. Ces clients nécessitent peut-être des stratégies de réengagement pour augmenter leur activité.

Partie 2 : Essais de différents approches de Modélisation

Interprétation des Centroïdes et des Descriptions de Clusters

Cluster 2 : **Clients à potentiel de croissance**

Recency : -0.033731 (presque neutre) -> Ces clients n'ont pas acheté récemment, mais pas il y a très longtemps non plus.

Frequency : 2.963620 (très élevée) -> Ils achètent très fréquemment.

Monetary : 2.491135 (très élevée) -> Ils dépensent beaucoup.

Interprétation : Ce cluster représente des clients extrêmement actifs, avec une fréquence d'achat et un montant dépensé bien supérieurs à la moyenne. Ce segment pourrait correspondre à nos meilleurs clients ou à ceux qui sont les plus engagés.

Partie 2 : Essais de différents approches de Modélisation

Déductions métiers

Stratégies Marketing Personnalisées :

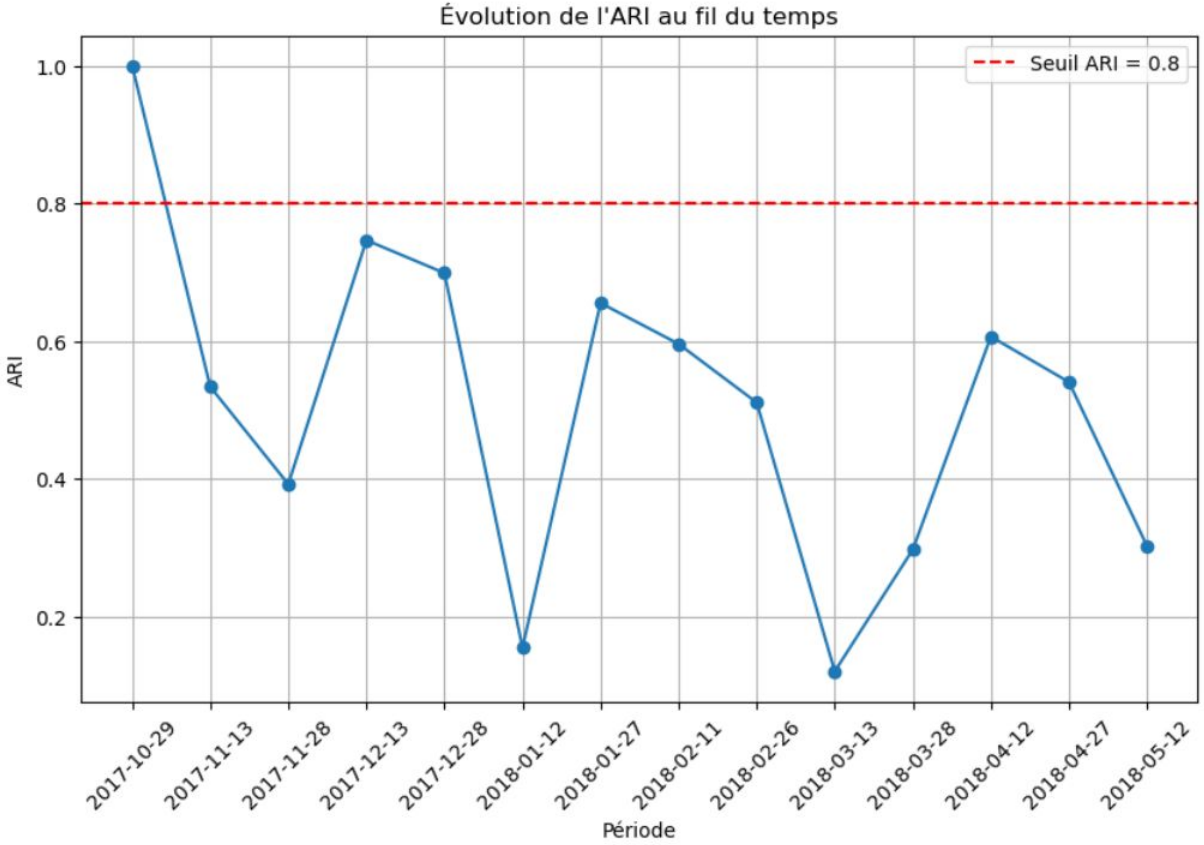
Clients Premium (récents et réguliers) : Ils pourraient être réceptifs à des offres exclusives ou à des programmes de fidélité pour renforcer leur engagement.

Clients à faible valeur (inactifs ou en désengagement) : Ils nécessitent des efforts de réactivation, comme des promotions ou des campagnes de réengagement.

Clients à potentiel de croissance (Meilleurs clients) : Ils peuvent être ciblés par des stratégies visant à maintenir ou à augmenter leur niveau d'engagement actuel, comme des offres VIP ou des récompenses pour fidélité.

Partie 3 : Simulation d’un contrat de maintenance

Partie 3 : **Simulation d'un contrat de maintenance**



Conclusion:

La segmentation doit être mise à jour tous les **15 jours**.

Conclusion

Conclusion

K-means est la meilleure méthode de segmentation des clients de ce site e-commerce.

La mise à jour du modèle de segmentation est à faire chaque **15 jours.**